Visual Attention Complexity of Scene

Qianpei Mai

L'Université de Nantes
Polytech Nantes

MASTER MDM
"Multimedia and Database Management"

Visual Attention Complexity of Scene

Author:
Qianpei Mai

Supervisor:
Patrick Le Callet
Emilie Bosc

August 22, 2013

Abstract

The visual attention complexity of a scene is regarded as a feature describing many information of the content in a video sequence, which meanwhile plays an important role in various applications to estimate the perceptual quality, information retrieval and so on.

In this paper, the hypothesis about visual attention complexity is proposed. The complex video sequence in the view of visual attention should contain a large quantity of "informative" objects on the scene. So, the visual attention complexity(VAC) indicator extraction from a video sequence is conducted by information theory on saliency map generated from computational visual attention model. The VAC indicator's performance is analyzed with the ground truth from an eyetrack database of IRCCyN/IVC.

The proposed VAC indicator is applied in video quality estimation methods which is widely used in video transmission or compression system. In addition to VAC indicator, spatial and temporal information from original video sequence and Bitrate or PSNR from compression video compose the set of elements for the quality estimation. The objective video quality estimation model is based on a machine learning algorithm and tested on a H.264 compressed video database of IRCCyN/IVC. All proposed models are verified by another H.264 compressed video database of IRCCyN/IVC. The quality scores predicted by proposed models have a high correlation coefficient to the subjective quality scores.

Contents

1 Introduction

1.1 Background

In these few decades, it has become the age of picture reading as people have to and prefer to receive information by pictures and video sequences. This phenomenon offered a new orientation of research in various fields. In the field of computer science, researchers were interested in digging out the information underneath the scene, describing the scene in different ways and so on. On the other hand, the biologists paid attention on the mechanism of human body (such as eye, brain and optic nerve) to perceive and process visual information. The study about how the content impacts on human gaze behaviors and how to improve the quality of experience when people watch the scene are the new areas of image processing combining with psychics.

According to Snodgrass in [SV80], visual complexity is one of four variable of central relevance to memory and cognitive processing to standardize scenes. In the term of pictures, visual complexity may affect such variables as naming latencies, tachistoscopic recognition thresholds, and memorability.

Visual complexity played an important role in various aspect. It can be a factor for designing the image or video objective quality metric, and also for building the quality experiment database [Win12], as visual complexity and quality rating are both related to psychological cognizance. Furthermore, visual complexity could be a clue to determine the ratio of compression and how to allocate the storage to memory the images or video sequences [FCLCJ12]. For example, details may be lost when visual complexity and compression ratio are high, or the visual complex region needs more bits to depict. Moreover, visual complexity is a feature for scene description and recognition, being used in content-based image retrieval and classification.

For the reasons mentioned above, many researchers have spent efforts to define and measure the visual complexity of images for the last few decades. Yet, the standard measurement have not validated until nowadays. Most of the proposed models just focused on estimating image complexity, not mentioned about video sequences. When it comes to video sequences, the complexity measurement must be influenced by some other factors, so image complexity metrics are not suitable. In addition, most of the metrics rely

on the patterns or structure of scene only without taking visual attention information into account.

1.2 Motivation

In this thesis, we attempted to estimate the complexity of scene based on visual attention which is provided for video sequences. The new indicator aims at measuring the visual attention complexity of scene, according to the saliency map, generated from computational visual attention model, of each frames on the video sequence. The saliency map is predicting the summarized human fixation distribution on the frame. The visual attention complexity is derived from analyzing saliency map conducted by Information Theory.

In view of bandwidth of transmission channel and size of storage space, it is required that video sequences are compressed to a certain size more or less. However, in the meantime, the quality of video should be controlled in a certain level, which must not fall into a low degree dramatically. Consequently, the objective video quality assessment is needed. Given the results for subjective quality experiments, even though two videos containing different contents have the same bitrate or peak signal-to-noise ratio, they are rated on different quality scores by human subjects. The visual attention complexity of scene proposed in this thesis as a measurement of the content's visual complexity is implemented to estimate video quality with bitrate or PSNR and other perceptual information. The quality prediction model is trained by machine learning algorithm with the data collected from subjective quality experiments.

These video quality assessment models can be applied to the video compression system. Before compressing, it is able to find out an optimal compression level, which would not cause a dramatical drop of video quality by analyzing the original video sequence.

1.3 Overview of the thesis

This thesis is divided into five sections. The first section is introducing the background of the topic, the motivation of our work and the outline of this thesis. The second section provides the overview of state-of-the-art

about visual attention complexity both on image and video. Meanwhile, the overview of objective video quality evaluation methods are provided especially perception-oriented ones, as visual attention complexity is a kind of perceptual information. In section 3, we presente a hypothesis of visual attention complexity of scene, the review of computational visual attention model, the extraction method of VAC indicator and experiments. The forth section introduces our video quality evaluation model with visual attention complexity, as well as experiments and discussion. In the final section, we make a conclusion of this thesis and point out the future work.

2 State-of-the-Art

2.1 Visual Attention Complexity

Complexity has become an active area to be studied in many different fields, such as information theory, computer science and psychics. The definition of a complex object in Webster's dictionary(1986) is to be "*an arrangement of parts, so intricate as to be hard to understand or deal with*". According to W. Li's, the quantity of complexity should be close to certain measures of difficulty concerning the construction, description of an object or a system.[RFS05]

Although the research of complexity in different fields defined many different measures, the questions which were asked by the researchers in each field about the complexity of their different subjects can be grouped into the same questions. These questions are "*How hard is it to describe?*", "*How hard is it to create?*", "*What is its degree of organization?*". [Llo02] So, the answers for measuring the complexity are considerable similar to each other regardless the fields.

However, the research of visual attention complexity dose not just relate to only one field. It is relevant to computer science and cognitive science. [DCE11] So far, the definition of visual attention complexity is nonspecific. It needs to be explored more in details.

2.1.1 Image visual attention complexity

Snodgrass and Vanderwart [SV80] regarded visual complexity of image as the quantity of detail or intricacy of line in the picture. They designed a subjective experiment with black outline drawings pictures asking observers to identify the name, judge the familiarity, rate the complexity degree and judge how similar the mental image of an object is to the pictures. They supposed the high visual complexity may lead to disagreement on naming the picture and unfamiliar. Refer to the result, visual complexity is negatively correlated to familiar. Yet, it has a positive relation with name agreement.

Forsythe et al. [FMS08] studied how the measures of complexity affected by familiarity with subjective experiments based on S&V 's work [SV80].

They found out that human rating complexity, as the normative metric, is influenced by familiarity. After, they explored four image-processing techniques, such as Perimeter detection, Canny detection. Perimeter detection metric has a strong correlation with human rating of visual complexity of line drawings objects and nonsense shapes. Meanwhile, it also eliminates the familiarity-bias from subjective judgments. Perimeter detection is introduced as a contour-based shape description, simple to be implemented. Hence, it is a popular visual complexity measurement.

Spatial frequency information is found out related to visual complexity of icon [FSS03] [For09]. Subjects tend to rate high visual complexity when the icon contains a larger amount of high spatial frequency information.

Compression ratio of image file is another common method to quantify visual complexity. In [FMS08], the authors explored two type of compression, JPEG and GIF. JPEG is a lossless compression technique, which allows the original image reconstructed from the compressed one. GIF has a good performance on pictures with limited colorization under 245 and sharp transitions.

Unlikely the previous research, Matthieur et al. [DCE11] considered the perception information while designing the visual complexity measurements. The heatmaps from a computational model of attention tend to have different spreading and patterns depending on image visual complexity, Figure 1. The authors applied JPEG compression on eye-tracking heatmaps, saliency maps and heatmaps from visual attention model and measure image visual complexity with compression ratio. They also calculate the saccade length Fourier entropy to judge whether there are temporal redundancies in the evolution of distance between two consecutive focus points.

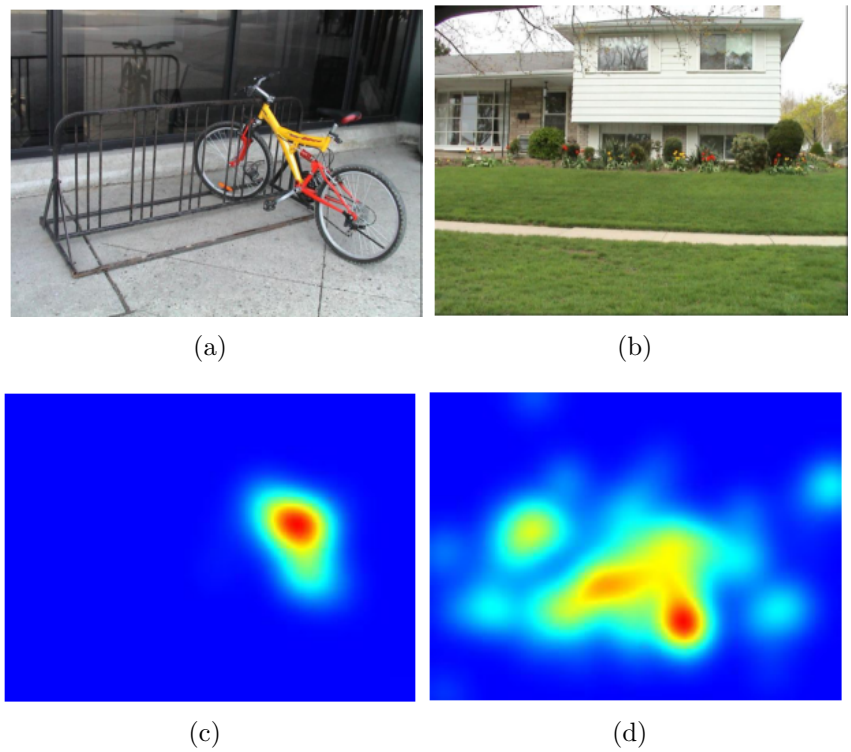


Figure 1: The subjective evaluation of visual complexity is a difficult task, Is (a) more complex than (b)? This question may be partially answered by observing their corresponding attention maps (c) and (d).[DCE11]

2.1.2 Video visual attention complexity

Even though there are plenty of methods to quantify the image visual attention complexity, the existing methods only considered the spatial information. When it comes to video visual attention complexity, temporal information should not be ignored. Hence, an appropriate extension to the temporal domain should be considered instead of applying directly an image visual complexity metric to video sequences.

Visual attention deviation (VAD) proposed by Feng et al. [FCLCJ12] is a measurement of video visual attention complexity. VAD determines how busy a video is. Feng referred the "busy" video to frequent shifts of visual attention when observer is watching the video, such as a music video. A

"quite" video induces few shifts of visual attention, such as a head-and-shoulders presidential address. The speedy shift of visual attention from on position to another in the spatial domain, is one of the typical eye movement – saccade. The other two typical movements, contrasting to saccade, are fixation and pursuit.

Feng et al. [FCTJ11][FCLCJ12] introduced a Hidden Markov Mode (HMM) for predicting eye movements along a video sequence using saliency map. Hidden Markov model can be considered as the simplest dynamic Bayesian network. The latent states of HMM are three types of eye movement, i.e. Fixation, Pursuit and Saccade, which could not be observed directly. The variable X_t is the hidden state at time t, where $X_t \in \{F, P, S\}$. The possible observations of HMM are all positions of pixel in the frame, which would be the output of the model. In time t, observation Y_t is corresponding to highest possible latent state. As human gazing behavior is different in these latent states, three schemes are proposed to predict Y_{t+1}. For fixation, gaze will stay in a certain position with few vibration. The emitted position is $Y_{n+1} = Y_n + W_F$, where W_F is a white Gaussian random variable with variance σ_F^2. For pursuit, gaze may follow a certain object from time t to time $t+1$. Then, the observation is $Y_{n+1} = Y_n + v_n + W_P$, where v_n is the pixel velocity vector and W_P is a white Gaussian random variable with sigma σ_P^2. For shifting gaze position – saccade, it is known as the fastest movement. Feng establish a gaze vector $g_{n-k:n}$ to predict the observation. The gaze position in time $t+1$ is $Y_{n+1} = Y_n + g_{n-k:n} + W_{s,k}$. In fact, $g_{n-k:n}$ could not measure during saccade, so, $g_{n-k:n} + W_{s,k}$ is replaced by W_G with a fairly large sigma.

The transition probability $\alpha_{i,j} = P(X_{n+1} = j | X_n = i)$ of HMM is derived from saliency map of frame. In order to determine the highest possible latent state, Feng employed forward algorithm to compute the latent probability.

$$\begin{aligned} &P(X_n = j) = \textstyle\sum_i P(X_{n-1} = i)\alpha_{i,j} P(Y_n | Y_{n-1}, X_n = j) \\ &P(X_0 = i) = \pi_i \end{aligned} \tag{1}$$

where $i, j \in \{F, P, S\}$. Then, the steady state probability of saccade states for the video is calculated from the models, which is defined as the VAD. Feng verified the VAD matching the ground true probability of saccade, resulting in measuring the degree of busyness for the video.

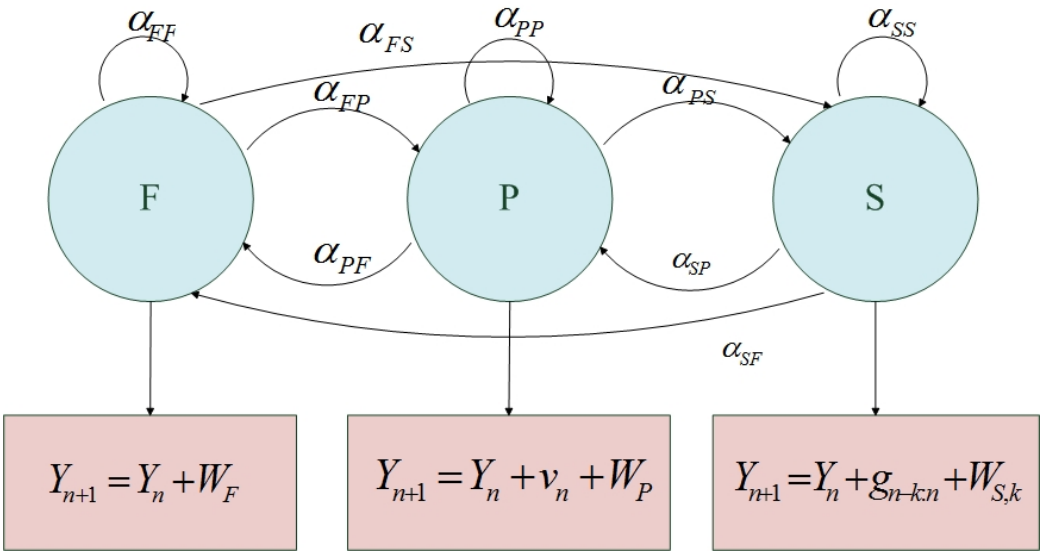


Figure 2: Hidden Markov Model of eye gaze during video observation. Circles denote latent states of F (fixation), P (pursuit) and S (saccade). Boxes denote observations.[FCTJ11]

To sum up, the advantages of VAD to quantify the visual attention complexity of video are: (i) The model is generated based on saliency information. (ii) It takes the unique characteristics of video into account, as it is a bottom-up model and consider the temporal clue. (iii) It is demonstrated close to the ground truth eye-track data. However, the drawback is that the transition matrix of this HMM should be exclusively derived for different contents of video, which reduces the operability.

2.2 Perception-oriented Video Quality Evaluation

As the digital video is widely used around the world, it is essential to measure the quality of video for various purposes. The video quality rated by people is the most convincing metric, obtained from Subjective Video Quality Measurements. The main idea of subjective quality measurement [ITU08] is estimating the average viewer opinion on the quality of one video. Yet, carrying out the subjective experiments is quite expensive in terms of time and human resources, also in terms of financial budget.

Objective video quality evaluation aims at approximating the mean opinion score(MOS) of subjective quality assessment.There are three basic types of objective video assessment metrics: (i) Full Reference Method (FR), that compares the difference between frames from distorted video and their corresponding frame of original video directly; (ii) Reduced Reference Method (RF), that compares the features extracted from original video and distorted video respectively, when all the original videos are not available for users; (iii) No Reference Method (NF), that derives a quality score just by analyzing the distorted videos without any reference to original one. The benchmark of objective video quality assessment is given by Video Quality Experts Group (VQEG).

Numerous objective video quality metrics exist [Win07], such as PSNR, MSE and so on. They are able to predict image degradation quite successfully and simple to implemented. However, they do not take the content and human gaze behavior into account. Sometimes, their good performance is related to particular types of distortion only [Win07].

In order to enhance the performance of objective video quality metrics, researchers studied the human perceptual visual system and purposed the perception-oriented video quality assessment. Two factors are most studied to improve the metrics, which are regions of interest and perceptual characteristics.

2.2.1 Region-of-interest-oriented video quality assessment

Because the gaze of a viewer has a large probability of falling on the region of interest when watching a video, the quality in the region of interest may have a greater influence on the global subjective quality of entire video sequence. Three kinds of region are investigated, that are particular region (e.g. face, edge), motion tune and saliency region.

For particular regions, in [CW04], a semantic segmentation driven metric is introduced. Péchard [PLCC+07] used the proportion of smooth areas, textured areas and edges as factors on bitrate repartition over the distortion sequence.

For the motion tune, Motion-based Video Integrity Evaluation index (MOVIE) [SB10] combined explicit motion information with visual quality assessment by the method of tracking perceptually relevant distortions along motion trajectories along spatial-time clue, leading to augmenting the measurement of spatial artifacts in videos. MOVIE measures both spatial and temporal video distortions over multiple scales, which is specific to video quality assessment.

When it refers to saliency map, the perceptual-oriented video quality metrics are refined by changing the weighting of local quality. In [FLYW08], Focus of Attention (FOA) is obtained from Itti's saliency map [IKN98] on distorted frame. Errors measurement is conducted between reference frame and distorted frame. The final quality score are computed after pixel-wise error pooling of FOA or saliency map and error map.

Certainly, some approach determines the visual attention region by several different factors. A visual Important Map judged by a number of factors such as shape, color, motion is defined by Osberger et al. [OBM98] The perceptual distortion map is weighted by corresponding important map before summation to a IM-weighted Perceptual Quality Rating. In [YKP10], the video quality is the average quality in the visual attention region i.e. saliency, motion and contrast of the distortion sequence.

2.2.2 Perceptual-characteristics-oriented video quality assessment

Since more and more research of perceptual characteristic carried out in the field of cognitive psychology and biology, objective video quality evaluation methods were designed guided by perceptual characteristics.

In [NLMLCB09], a perceptual full reference video quality assessment based on temporal evolution of spatial distortion is proposed. The first two steps of the method compute perceptual distortion maps for each frame with Wavelet-based Quality Assessment metric and processing motion estimation along the video. Temporal quality evaluation is done in the two following steps as short-term pooling and long-term pooling. In third step, a spatio-temporal tube is created in the fixation level, which may last for the duration of fixation. The mean distortion and variation in temporal level is calculated. In the

case that mean distortion is perceptible only when the variation of distortion is smoothed, the spatial distortion in the tube is filtered in temporal level. Long-term pooling are according to the evaluation process of human to an entire video i.e. "*quick to criticize and slow to forgive*". Not only the mean distortion of the whole sequence but also the temporal vatiation of distortions over the whole sequence contribute to the final global objective quality score.

Wang et al. [WSB03] presented a generic error-sensitivity based quality assessment which is based on Human Visual System modeling. Similar to the Daly's HVS model[Dal92], the framework of this method contains preprocessing stage, CSF filtering stage and channel decomposition. After these three steps, which aim at mimic human visual system, error normalization and masking is processed on each channel with a visibility threshold model. Following, the errors of all channels are combined into a single value to estimate the objective quality of video.

2.3 Conclusion

In this section, the overview of existing visual complexity estimators both on image and on video has been presented. They have both shown to make contribution in different aspects of application. However, the visual complexity of image is not sufficient for video sequences. And, the visual attention deviation for video should derive uniquely for each content. We tend to design a visual complexity of scene indicator which is according to the saliency information and efficient regardless to the different contents.

On the other hand, this section contains the overview of objective video quality assessment methods, especially the perception-oriented ones. The perceptual information is verified to improve the performance of objective video quality estimation. In out work, we want to design a new method of quality assessment with visual attention complexity.

3 Visual Attention Complexity Based on Saliency Information

3.1 Hypothesis

According to Tsotsos et al.[TCKW+95], visual attention mechanism involves to four basic components: (i) the selection of a region of interest in the visual filed; (ii) the selection of feature dimensions and values of interest; (iii) the control of information flow through the network of neurons that constitutes the visual system; (iv) the shifting from one selected region to the next in time. These cognitive processings impact on the eye movements of human.

There are several types of eye movement, but the two most basic and interesting are "fixation" and "saccade". Fixation is the visual gaze maintaining on a certain location, which can last on ranging from 100 ms to 600ms.[RP92] Saccade is high-speed movements from one object to a different part of the same object or to another object, which only lasts approximately between 150 and 200 ms.[Pal99] Therefore, people tend to fixate on regions of interest, especially the semantically "informative" objects.[CBD02] Saccade would appear at the end of a fixation and be followed by a new fixation changing the visual attention on new object.

In order to describe complexity of visual attention regarding video sequences, a hypothesis is proposed that in a visual attention complex video, there is a larger quantity of "informative" objects on the scene. In other words, an observer would fixate on several regions of interest along a video sequence, or a group of observers would fixate on different regions of interest in the same scene. On the contrary, while watching a video with a small amount of "informative" objects which is not complex in terms of visual attention, an observer may continue fixating on a certain object, including in the case of pursuing a moving object. However, an exceptional case exists. If the video has nothing interesting, people may fixate on every part of the scene with same statistical possibility, at the same time, changing visual attention region as soon as they complete processing the region. Hence, these "monotonous" video sequences are categorized into visual attention complex ones.

The method of measuring visual attention complexity of scene according to the hypothesis above is required. Two properties of eye movements, "*Where the movements are made to*" (spatial measures) and "*When the movements are made*" (temporal measures), are patterns to gain insights into cognitive processes.[CBD02] Eye movements recording by eye tracker are the precise information for the research of visual attention complexity. Even so, in general, saliency information provided by visual attention models are comparatively simple and practical to use widely. In consequence, in th rest parts of this section, a visual attention complexity indicator based on saliency information is proposed.

3.2 Visual Attention Models

3.2.1 What is visual attention models?

The computational visual attention models aim at detecting the location that attract the gaze of an observer. Most of them provide a saliency map indicating the position of most visual interesting parts .[MC09] Saliency map computation model is based on Feature-Integration Theory of Attention[TG80]. According to Tsotsos et al.[TR11], the first model in this class is a model by Koch[KU87]. Five elements are included: (i) computing a set of feature maps, permitting represent several stimulus characteristics separately; (ii) encoding a topographic saliency map be combination properties across all feature maps; (iii) selective mapping into a central non-topographic representation, through the topographic saliency map; (iv) a winner-take-all (WTA) network implementing the selection process based on conspicuity of location; (v) inhibition of this selected location causing an automatic shift to the next most conspicuous location.

Saliency map is a combination of information from several feature maps. Computational visual attention model determine a saliency region considering the degree of difference between that location and its surrounding.[TR11] The most classical visual attention model, which is usually regarded as the benchmark being compared the performance with new models, is proposed by Itti[IKN98] which is a purely bottom-up model building on Feature integration theory and framework of Koch's model[KU87].

Itti's model is a simple computational model, which can efficiently detect saliency region depending on attention-focusing features (colors, intensity and orientations). Whereas, this model is not able to detect the saliency region for unimplemented feature types. Moreover, when it comes to video, pixel value different from frame by frame is not considered by this model.

3.2.2 Graph-Based Visual Saliency

Graph-Based Visual Saliency (GBVS) is also a bottom-up visual saliency model, introduced by Harel et al.[HKP06] Although it has the similar structure as other hierarchical models, due to graph computations, it is more powerful to predict human fixation against benchmark and Itti's model.

GBVS are organized into three stages as well: (i) biologically motivated feature extraction; (ii) activation maps related to each feature channel generation; (iii) normalization activation maps to highlight conspicuity and combination them into a final saliency map. The contribution, that is the difference with others, is implementing graph algorithms in second and third stages. Typically, subtracting feature maps are applied in second stage, and the third stage employs normalization on local maximum, a difference-of-gaussians filter or a nonlinear interactions. GBVS accomplishes two Markov chains to achieve goals of second and third stages.

In feature extraction stage, GBVS produces different biologically inspired filters computing feature maps, including some normal features(e.g. intensity, orientation and etc.) and contextual features. There are two contextual features in stage one, flicker and motion. Flicker is the absolute difference between previous frame and current frame. As it is mentioned in [CBD02], the gist of scene would be the region people may fixate on it, even though it is "semantically uninformative". Motion feature are detected between previous frame and current frame along four directions 0°, 45°, 90°and 135°.

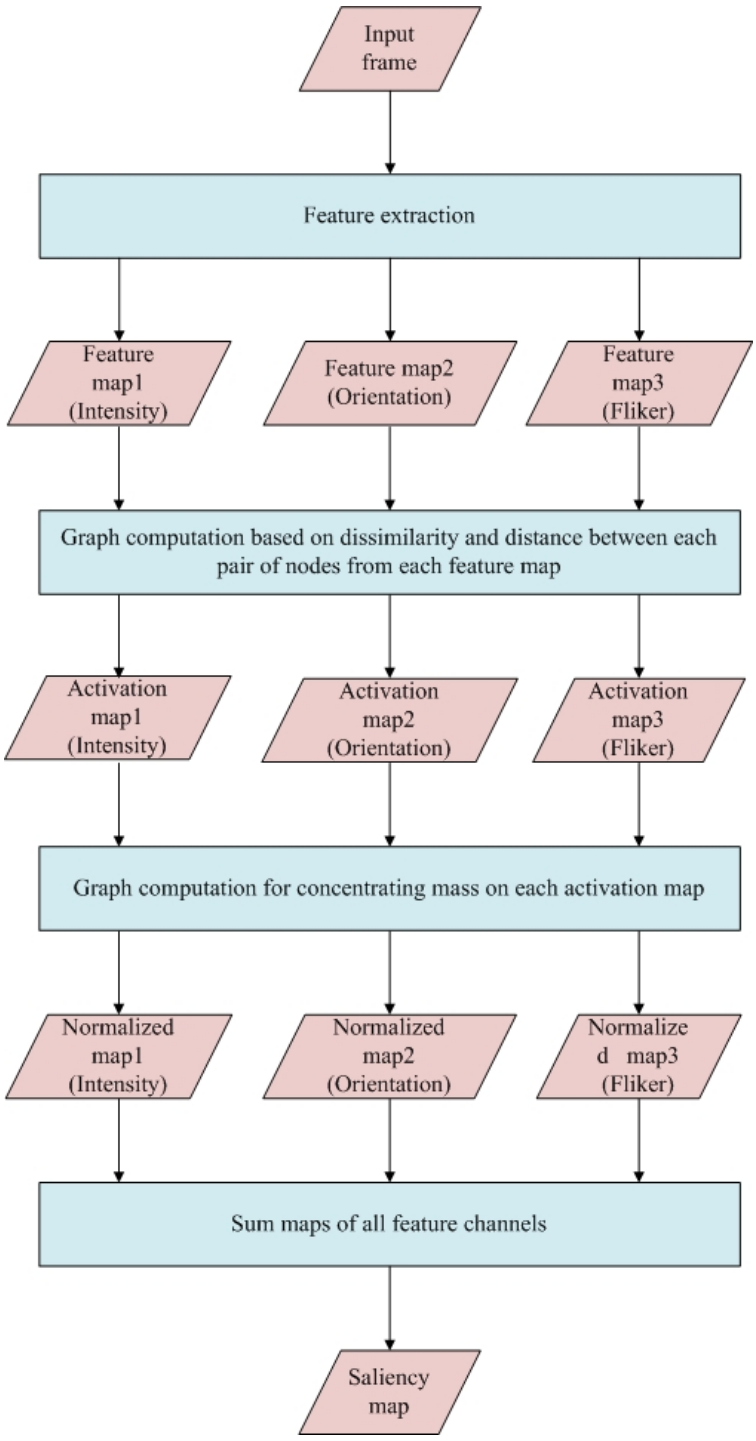


Figure 3: Flow-chart of GBVS

In second stage, the model forms activation maps based on Markovian approach. An activation map corresponds to a certain feature map, which position with a high values represents more unusual to its surrounding. Providing that a feature map M is $n \times n$ dimension, a fully-connected directed graph G_A with n^2 nodes is generated related to every position of M. From one node (i, j) to another node (p, q), the directed edge is determined by dissimilarity and distance between these two positions in feature map M, seeing function(2). Definitely, the weight of opposite direction in this edge is the same. In order to generate a Markov chain, the total outgoing and incoming weight should be normalized to 1. The normalized weights draw the transition probabilities of the chain. If a node is dissimilar with its surrounding nodes, there would be a mass in that node accumulated by equilibrium distribution, since it has higher transition probabilities.

$$\omega_a((i,j),(p,q)) = d((i,j) \parallel (p,q)) \cdot F(i-p, j-q), \tag{2}$$

where

$$d((i,j) \parallel (p,q)) = \left| \log \frac{M(i,j)}{M(p,q)} \right|, \tag{3}$$

and

$$F(a,b) = \exp\left(- \frac{a^2 + n^2}{2\sigma^2} \right). \tag{4}$$

In third stage, normalization stage, a n^2 nodes graph G_N is for normalizing an activation maps. Edge weights between each pair of node (i, j) and node (p, q) in G_N are corresponding to activation value of end node and distance between them, seeing function(5). According to G_N, a new Markov Chain is capable of locating the concentrating mass. Comparing to three methods of normalization mentioned above, this concentrating mass an activation maps determines a few but important key saliency regions, instead of a nearly uniform result.

$$\omega_n((i,j),(p,q)) = A((i,j) \parallel (p,q)) \cdot F(i-p, j-q). \tag{5}$$

3.2.3 Why GBVS is chosen?

There are five main reasons why we choose GBVS to extract saliency information of videos in our work: (i) Contextual information is considered,

which is an unique characteristic of video sequences; (ii) Center-bias is observed on the saliency maps, which also be observed on real-world eye movements recording[Tat07]. Because the node in central area are average closer to other points in the images, the equilibrium distribution would assign a higher value of central nodes. (iii) Unlike some visual attention models, the saliency map provided by GBVS highlights a few key region, rather than a nearly uniform map. (iv) The author implemented a multiresolution version in activation and normalization stage to improve the performance. (v) Even if GBVS processes two Markov Chains and multiresolution of maps data, this model can be computed parallelized and efficiently.

3.3 A VAC indicator based on saliency map entropy

To answer the first question for complexity study summarized by Seth Lloyd[Llo01] – "*How hard is it to describe?*", Seth listed plenty of metrics, containing entropy, which would be one of the most widely applied measurement.

In information theory, entropy is a measure of information content describing uncertainty, surprise and randomness. Typically, Shannon entropy quantifies the unexpected value of information contained in a content. Shannon introduced the definition of entropy H [Iha93], as followed. For a set of discrete random variable X having values $\{a_1, a_2, ..., a_m\}$ with relative probability $P(X = a_i) = p_i, i = 1, ..., m.$, the Shannon entropy of random variable X is

$$H(X) \equiv H(p_1, p_2, ..., p_m) = -\sum_{i=1}^{m} p_i \log p_i \tag{6}$$

In accordance with the hypothesis of visual attention complexity of video, the measurement of this type of complexity could be measuring how much is the quantity of "informative" pixels on the scene. A large area of "informative" objects leads to a complex scene. Since the values of each pixel in a saliency map represents how much is the probability an observer may gaze on it, "*Where people may fixate?*" would be obtained from saliency map, meaning that the high value in saliency map represents the high possibility this region will be gazed. In condition of a complex scene, a majority of pixels would have a large range of higher values. On the other hand, if there

is only one small region would be fixated in the scene, except extreme high values in these few pixels, other pixels just have values nearly 0.

We propose an indicator of visual attention complexity, which is the entropy of saliency map for a certain scene. The Shannon entropy of saliency map represents how "unpredictability" fixation regions appear on the scene. High entropy value denotes the scene has a lot of "informative" objects, so it is a complex scene in terms of visual attention. So, low entropy value indicates the scene is not complex in terms of visual attention.

In practice, in order to generate one visual attention complexity rating score for a video sequence, VAC is computed as below: firstly, calculate the entropy of saliency map for each frame; then, suppose the sequence has n frames, figure the median of these n entropy values as the final VAC.

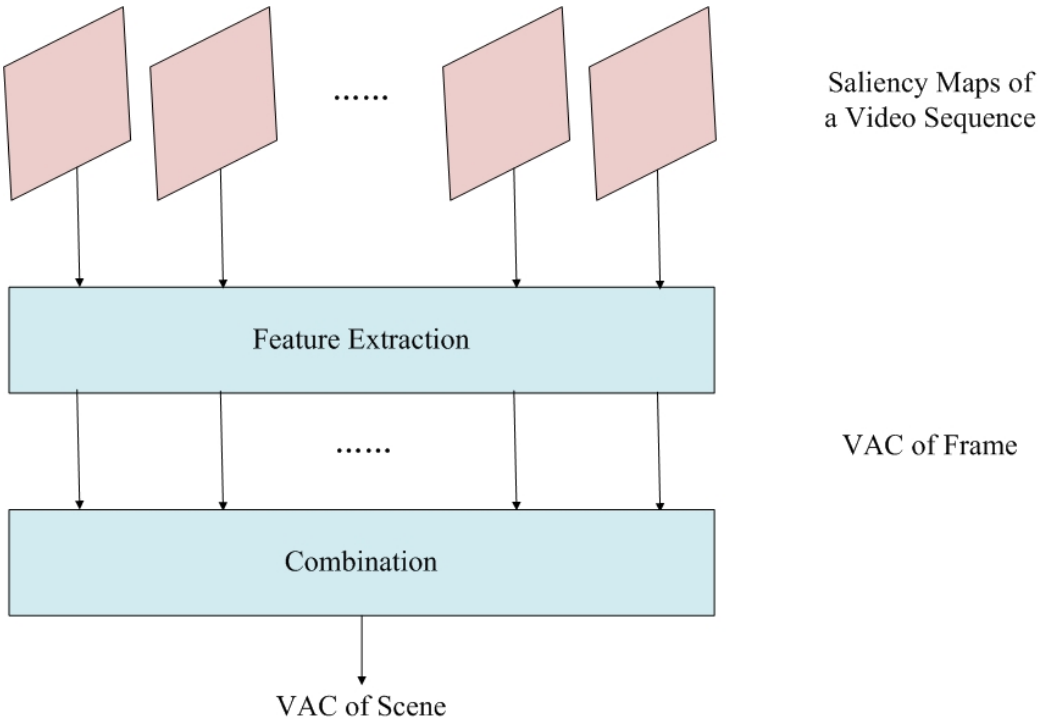


Figure 4: Flow-chart of VAC indicator based on saliency information

While extracting the VAC of frame, we just computed the entropy of histogram which has 256 bins $(0 - 255)$. In fact, feature can be extracted from percentile 75 of saliency map or other possible status. In addition, to combine n VAC values of n frames into one exactly value for a video, percentile 95 or average could be used.

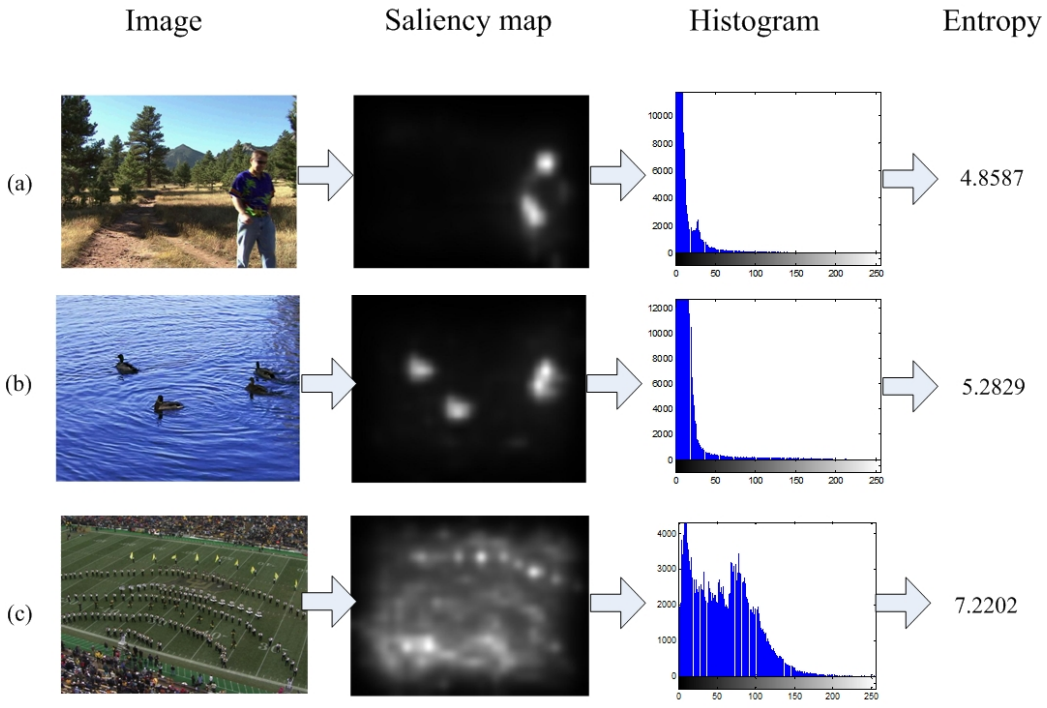


Figure 5: Illustrate how the VAC indicator measures the complexity of visual attention: (a)According to the saliency map, most of fixations fall in the man's face and his waving hand. So, pixels of saliency map are concentrated in a low value, which resulted in a low entropy. (b)Since there are four ducks in the scene, the regions of interest are more than scene(a). The entropy of saliency map is slightly higher. (c)As it is a scene of cheerleading squad performing, the region people may fixate is large. Depicted in histogram, values of saliency map are high in a wide range. The complexity of visual attention in this scene is high, which can be indicated by the entropy of scene.

3.4 Experiments and Results

3.4.1 Database

The VAC indicator performance is analyzed by the experiments with database IRCCyN/IVC_Eyetracker_SD_2008_11_Database [BCPLC09]. The first fifteen contents' original videos without compression or transmission errors are chosen to use in this experiments. The eyetrack data is from a free task experiment, meaning that the subjects were not asked to focus on any region. The recorded gaze positions reflected the nature attention of viewers. The eye gaze position is recored in 50Hz by monocular mode. So, in every record time, one position in the area of display is written down. The general condition of the subjective experiment is shown detailedly in Table 1.

Feature	Value
Resolution	SD(720×576)
DisplayMode	Interlaced
Frame rate	25
Number of observers	37 or 38 naives (depending the content)
ObservationDistance	6H
Luminance	$0.65/450 cd/m^2$
Duration	15 minutes
Eyetracker	Cambridge Research System EyeTracker
Eyetracker mode	Monocular
Eyetracker acquisition frequency	50 Hz

Table 1: General condition of IRCCyN/IVC_Eyetracker_SD_2008_11 database.

3.4.2 Experiment

According to Feng et al. [FCLCJ12], the proportion of saccade type eye movement during the whole video instead of fixation or pursuit could represent the visual attention complexity of the video, which is called as Video Attention Deviation by Feng.

The proportion of saccade in the eyetrack data of a video is regarded as the ground truth to compare with proposed VAC indicator. Since the

record gaze position from the database contains saccade, fixation and pursuit, we implemented the algorithm in Le Meur's previous work [LMNLCB10] to filter the fixation and pursuit movements. The algorithm is shown as below: (i) Calculate point-to-point velocity for each recored position; (ii) Label each sample below a given velocity threshold ($25°/s$) as belonging to a potential visual fixation period, otherwise as to a saccade period; (iii) Merge the group of fixation. According to the fixation duration must be higher than $100ms$, if the length of group is less than this threshold, the samples would be considered as saccade.

To calculate the proposed VAC indicator on the database, the saliency map for each frame in the video is generated by GBVS. Firstly, the frame is resized into $32pixels \times 32pixels$. The feature channels of the model are intensity and flicker. GBVS Multiresolution [HKP06] is uesd in this experiment, so as to represent feature maps by three-level pyramid.

(a) (b) (c)

(d) (e) (f)

Figure 6: Subfigure (a), (b) and (c) are frame 41, 51 and 61 of "DucksTake-Off". Subfigure (e), (f) and (g) are frame 112, 142 and 189 of "GroupDisorder". The proportion of saccade and VAC value of "DucksTakeOff" and "GroupDisorder" are $[0.3392, 5.8920]$, $[0.4123, 6.2052]$

3.4.3 Result and discussion

The relation between VAC and proportion of saccade is supposed to be directly proportional. In other words, the video having higher VAC should meanwhile have a higher proportion of saccade, meaning that it is visual attention complex. The results are depicted in Figure 7. Their correlation coefficient is 0.3245, which does not meet the needs.

These results can be explained through the analysis of the results, video contents and proposed VAC indicator, the reasons could be sum up to two points: (i) The VAC indicator only considers the total size of area having certain value of possibility to be fixated. Such as content "DucksTakeOff" Figure 6, there are four ducks on the lake, and in the video, they flied away. The indicator can discover that four main regions to be gazed, but ignore that people may pursuit the flying trajectory of the duck which they fixate before. To address such problem, the number of saliency objects and how much probability it can be fixated or pursuit should be determined. (ii) The bottom-up visual attention model could not detect the "semantical" objects interested to top-down mechanism. For instance, in content "GroupDisorder" Figure 6, nearly ten dancers danced into the scene and left. Viewers are likely to fixate on their faces. And each dancers stayed in the scene less than $1-2$ seconds, leading to large proportion of saccade in the scene.

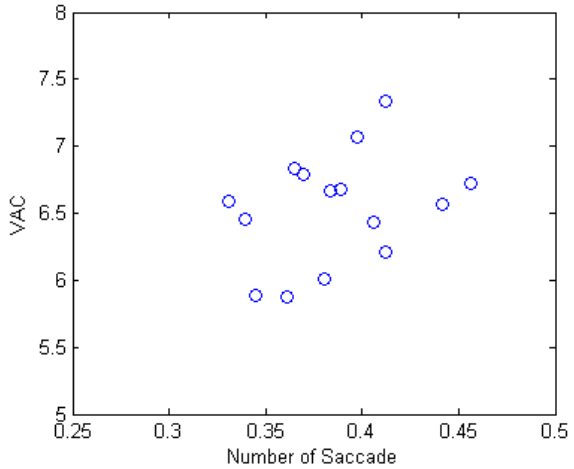


Figure 7: The proportion of saccade and proposed VAC of fifteen videos.

3.5 Conclusion

Because the visual attention complexity of scene is still ill-defined, we have to define the visual attention complexity of scene firstly. At the beginning of this section, we referred the video sequence which is complex in terms of visual attention to the video sequence which has a large amount of saliency objects. In the other words, in the case of a complex video sequence, in terms of visual attention, as we defined it, the viewer may fixate on several different regions.

In order to derive a VAC indicator, the knowledge of information theory is applied to precess the saliency information on the video. The saliency information is extracted by saliency map generated by the computational visual attention model. The second subsection discussed the visual attention models. Graph-based Visual Saliency model is employed because it is suitable for video and efficient. In the third subsection, the method to measure the visual attention complexity of scene is based on the computation of the entropy of saliency maps of the video sequence. The final VAC indicator is derived after the combination along the whole video sequence.

At the end of this section, the verification experiment has been presented. The ground truth is regarded as the saccade proportion by counting on the eyetrack data. Then, the correlation coefficient is computed between VAC indicator and ground truth. However, the result was not as good as expected. As explained in the section, it is because of the limitation of entropy and incomplete saliency information.

4 Quality Evaluation with Visual Attention Complexity

4.1 Motivation

4.1.1 Subjective Quality Evaluation

In general, the subjective video quality assessment is considered to be reliable. So, it would be the ground truth of quantifying the human perceptual quality of video sequence among of two types video quality assessments – "subjective" and "objective". The result of subjective assessment is mean score of opinion (MOS), which is the mean rating from a group of observers after watching the video sequences.

In order to obtain a result close to the real data as much as possible, also being able to use widely, the condition and methodology of subjective video quality assessment should strictly obeys to a standard. In 2008, International Telecommunication Union (ITU) instituted the standards of experiment environments and the categorized methodologies of subjective test in the Recommendation ITU-T P.910[ITU08].

In addition to the fact that the source signal should be chosen according to the goal of the test, the test environment needs to be taken into account. The test environment may seriously impact on the test results, such as lighting conditions, scene characteristics, viewing distance and so on. For example, the variation of light caused by AC frequency may lead to a flicker in the video sequence, or different viewing distance of the same video may result in different quality scores.

In Recommendation ITU-T p.910, they presented four test methods: Absolute category rating (ACR), Absolute category rating with hidden reference (ACR-HR), Degradation category rating (DCR) and Pair comparison method (PC). For the reason that this paper aims at predicting MOS from ACR-HR test, we will introduce more details of this method.

As ACR-HR is called "hidden reference", reference version is contained in test sequence set and is rated as video with stimulus. In the data analysis

part, a differential quality score (DMOS) is computed between reference video and stimulus video. Five levels of rating score can be chosen by observers: 5 – Excellent, 4 – Good, 3 – Fair, 2 – Poor and 1 –Bad.

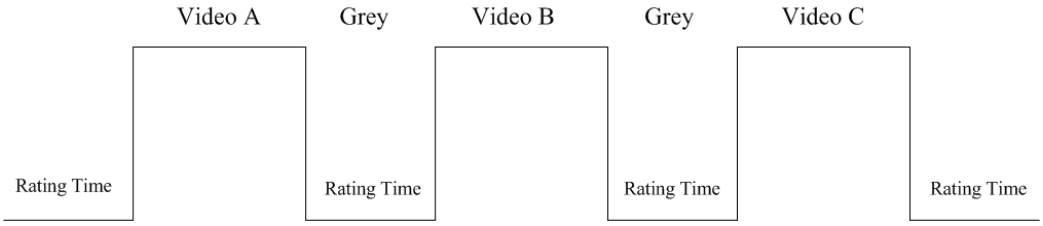


Figure 8: Stinulus presentation in the ACR-HR method. Video presented to observers could be compressed sequence or reference sequence without noticing to observer. Rating time follows by presenting time, which should be less than or equal to 10s.

Subjective video quality assessment is a benchmark for video quality research, but the drawbacks are extremely time consuming and unpractical to qualify each video sequence by a group of observers (regardless volunteer or get paid). Consequently, researchers spent effort on evaluating video quality which would be more relevant to subjective quality result.

4.1.2 MOS evolution with Bitrate and PSNR

One possible application from the outcome of this thesis is the quality evaluation of H.264-coded video sequences. H.264 developed by ITU-T Video Coding Experts Group (VCEG) is a standard for video compression, which is most widely used formats for compression nowadays. Obviously, compressed video sequences with different bitrates have dissimilar perceptual qualities. Generally, the higher the bitrate, the higher the MOS from subjective test.

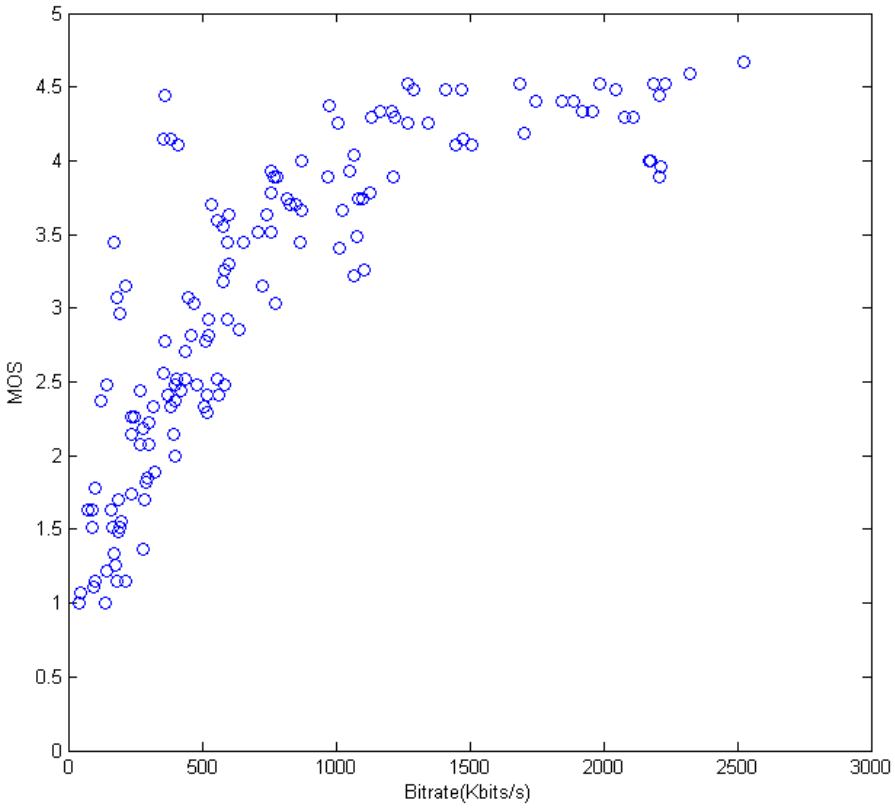


Figure 9: MOS evolution with Bitrate

The peak signal-to-noise ration (PSNR) is a kind of objective metrics being used in video community for a long time. In the field of video quality assessment, PSNR is known to have an approximate relationship with the quality perceived by human observers, successfully predicting subjective rating for some compression distortion.[Win07]

$$PSNR = 10 \log_{10} \frac{255^2}{MSE}, \; MSE = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2 \tag{7}$$

where N is the number of pixels,
$\varepsilon_i^2 = (x_i - x_i')^2$ is the square difference
between reference image x_i and distorted image x_i'.

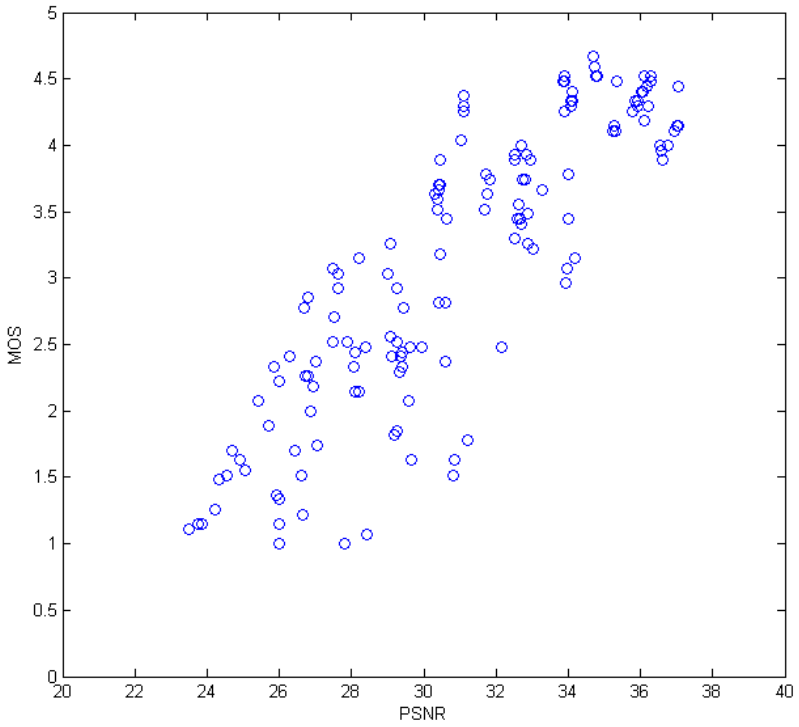
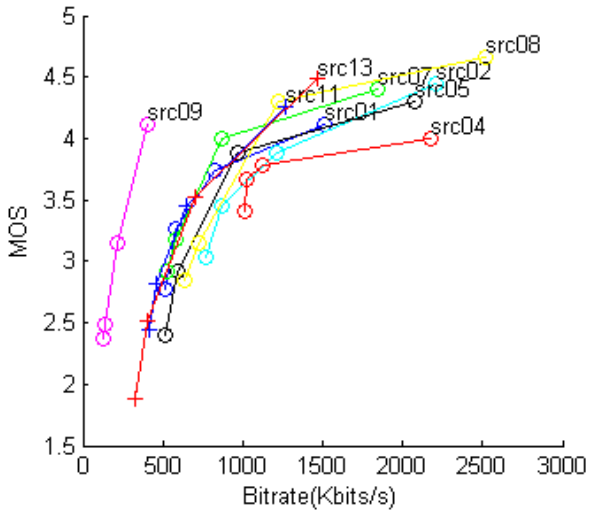


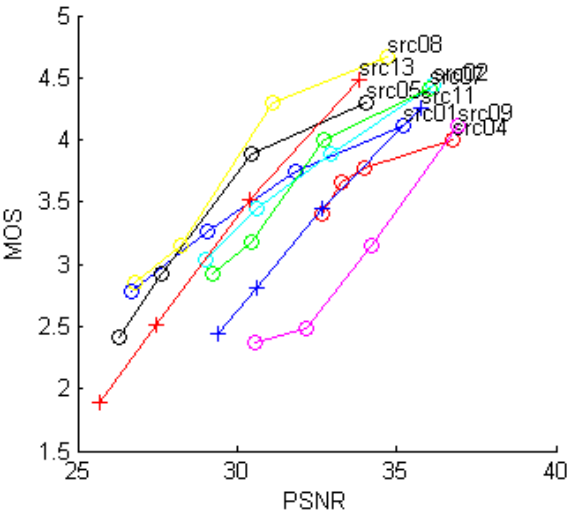
Figure 10: MOS evolution with PSNR

Roughly, Bit Rate and PSNR both have a direct proportion relationship with MOS, seeing Figure 9 and Figure 10. The relationship between MOS and two factors are obvious. The MOS increases in the logarithmic way leading by the increase of bitrate. And MOS and PSNR has a linear relationship, which is not tight enough. Whereas two figures below Figure 11, when it comes to each unique video, the shape of curve are unique as well, for instance, higher slopes of some video's curves, or the displacement of curves. The problem of bitrate and PSNR is they do not take the content into account, reflecting the fidelity of distorted video without characteristic

of the video itself.



(a) MOS–Bitrate of QP26



(b) MOS–PSNR of QP26

Figure 11: MOS–Bitrate and MOS–PSNR curves from videos compressed with QP26 in layer0 and different QP in layer1.

4.1.3 Content impact on the MOS prediction

The content of video sequence impacts on the evaluation of subjective video quality assessment. Consequently, features describing attribute of scene would give a contribution to MOS prediction.

Generally, perceptual information of content is a basic choice. The spatial and temporal perceptual information of the scenes are defined in ITU-T P.910[ITU08], as the critical parameters for compression of video.The spatial perceptual information (SI) aims at measuring the complexity of spatial detail in the scene, based on Sobel Filter, seeing Function (8).The temporal perceptual information (TI) assesses the quantity of motion based on pixel values along the video sequence, seeing Function (9). To be mentioned, the change of scene is concerned in temporal information.

$$SI = max_{time}(std_{space}(sobel(F_n))), \tag{8}$$

where F_n is frame in time n, n is from 1 to N,
std_{space} is standard deviation on soble-filtered frame,
max_{time} is maximum along the time.

$$TI = max_{time}(std_{space}(M_n(i,j))) \tag{9}$$

where $F_n(i,j)$ is (i,j) pixel of frame n,
$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$ is motion different feature,
std_{space} is standard deviation motion feature in each frame,
max_{time} is maximum along the time.

Despite of perceptual information, we tend to employ the feature representing visual attention information. So, visual attention complexity indicator of video (VAC) mentioned in the former section is a factor of the MOS prediction model introduced following.

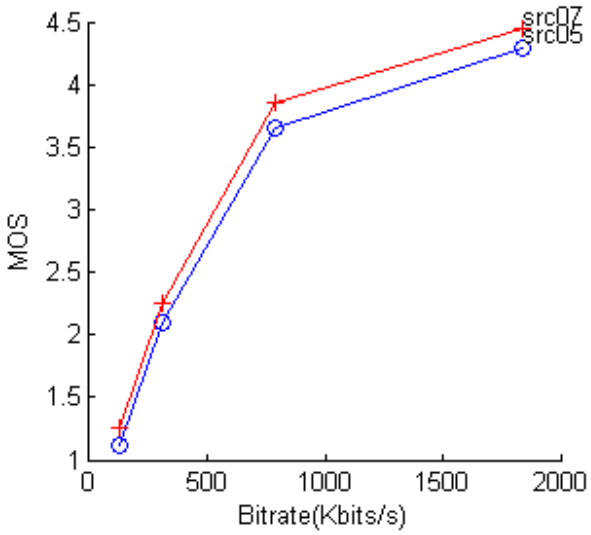


Figure 12: MOS–Bitrate curves of SRC05 and SRC07 have a resemble shape. Meanwhile, these two sources have similar values of SI, TI and VAC, which are $[139.09, 17.90, 6.60]$ and $[138.13, 19.72, 6.47]$.

4.2 MOS Prediction Models

In order to enhance the ability of bitrate and PSNR in the field of predicting MOS, we proposed a reduced-reference (RR) model combining a set of feature simply attained with video sequence itself. There are several models combine a set of feature. They analyzed the relationship between each feature and MOS, then formed a function combine the features and MOS. They applied Linear Regression[Roq09] or Partial Least Squares Regression[OD07] to generate the final model. However, we would generate our model in a new way. We firstly expose the relationship between the features and MOS by Symbolic Regression, and then apply Linear Regression to retrain the model so as to refine the fitness of the model to a certain database.

4.2.1 Model training

Training database

In order to train the basic MOS model with Symbolic Regression, nine out of the eleven video contents from IRCCyN/IVC_SVC4QoE_QP0_QP1_Video_VGA

database[PEB+11] are selected as the training set, Figure 13. There is no compression processing on reference sequences. 16 different Hypothetical Reference Circuit (HRC) , which are H.264/SVC, are coded with different QP in base layer and enhanced layer respectively.

No.	QP_Layer0	QP_Layer1	No.	QP_Layer0	QP_Layer1
HRC01	26	26	HRC09	38	26
HRC02	26	32	HRC10	38	32
HRC03	26	38	HRC11	38	38
HRC04	26	44	HRC12	38	44
HRC05	32	26	HRC13	44	26
HRC06	32	32	HRC14	44	32
HRC07	32	38	HRC15	44	38
HRC08	32	44	HRC16	44	44

Table 2: Compression paramerters – QP setting of each HRC.

The values of features in the training dataset should be variable and spread in a range. The reason why "Stream" and "Family" are eliminated from the dataset is because the Bit Rate of H.264-compressed sequence of these two contents is in the range of $(800, 11100)Kbits/s$ and $(100, 7000)Kbits/s$, considered to be the outliers, Figure 14. However, the other coded sequences' bitrates fall in the range of $(0, 2600)Kbits/s$. Except for bitrate, the other three features – SI, TI and VAC are extracted from reference video as the description of content.

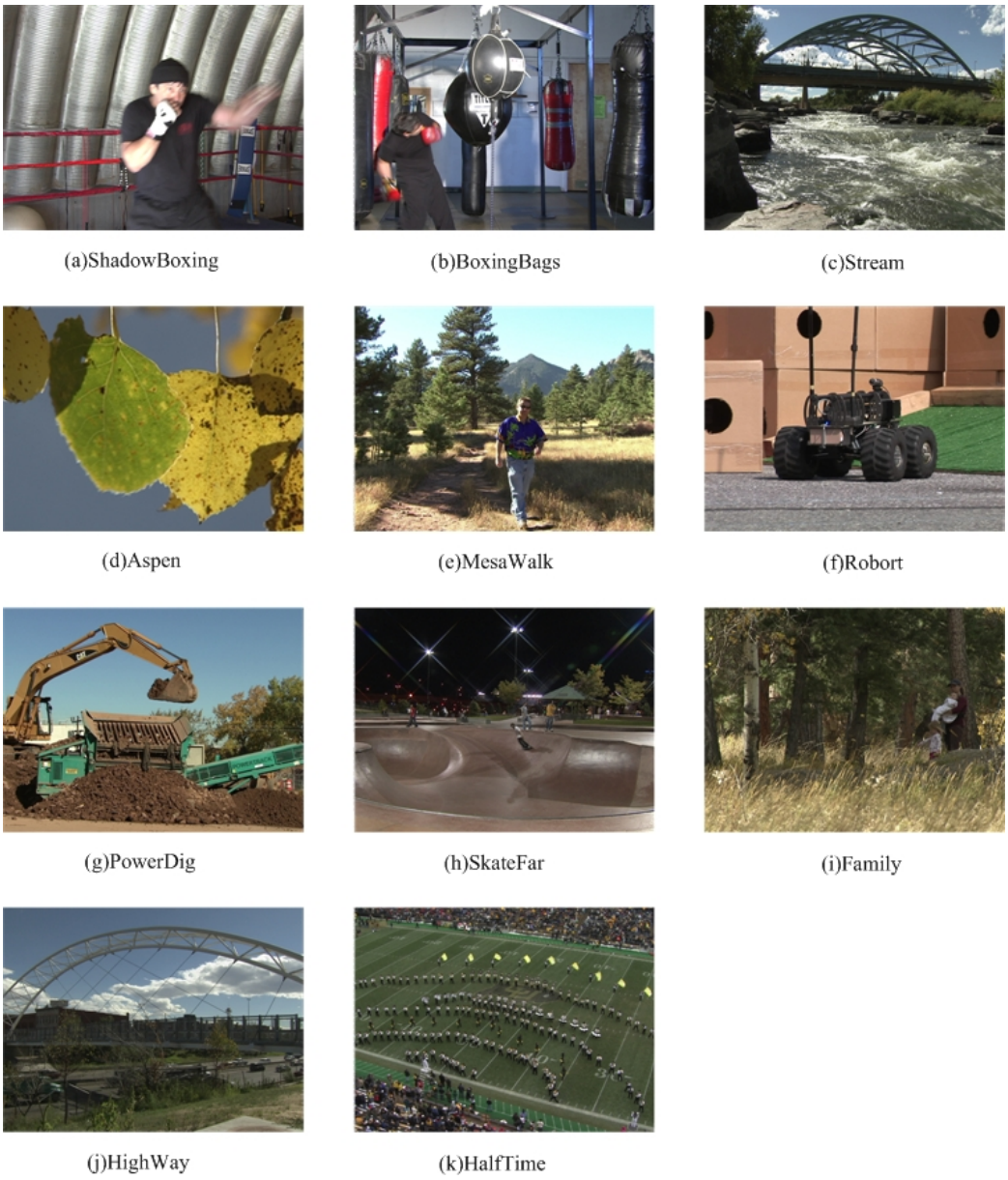


Figure 13: Frame shot from IRCCyN/IVC_SVC4QoE_QP0_QP1_Video_VGA database.

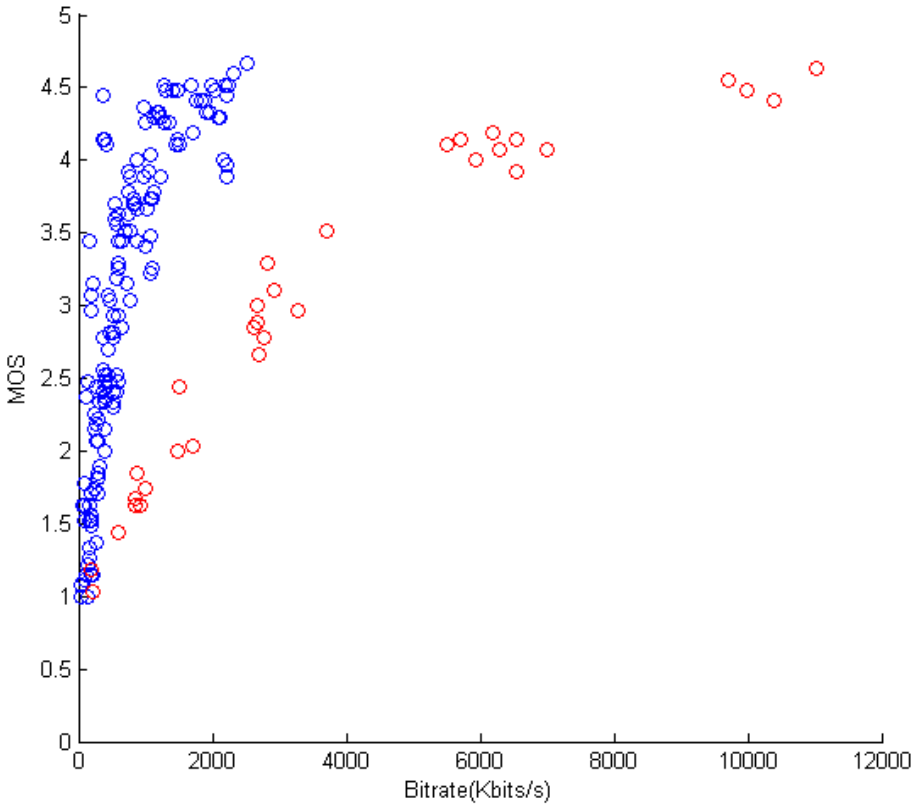


Figure 14: Red points are data from "Stream" and "Family", which obviously have much higher bitrates.

Subjective quality assessment experiments are done by 27 observers with the methodology – ACR-HR. MOS is computed from the test results in the experiment condition shown in Table 3.

Feature	Value
Resolution	VGA(640×480)
DisplayMode	Progressive
Number of observers	27
ObservationDistance	4H
Luminance	$0.1/180 cd/m^2$
Duration	$2 \times 50 minutes$

Table 3: General condition of IRCCyN/IVC_SVC4QoE_QP0_QP1_Video_VGA database.

Symbolic regression

Symbolic regression[SL09] is originated from evolutionary algorithms. It is used by Koza[Koz92] as genetic programming, the basic form of symbolic regression. Distinct from linear regression,the goal of symbolic regression is to derive the mathematical formula in order to fit a set of parameters, such as $y = f(x_1, x_2) = \log(x_1) + x_2^2 + C$, based on a set of arithmetic operator (e.g. +, sin, exp) even logical operator (e.g. If-Then-Else, Less-Than). Instead of one "optimal" solution, symbolic regression supplies multiple feasible candidates.

After collecting and computing all the variable (i.e. MOS, Bitrate, PSNR, SI, TI, VAC) from training video dataset, they are inputed into the algorithm. The target expected formulas are set as

$$MOS = f(Bitrate, SI, TI, VAC) \tag{10}$$

$$MOS = f(PSNR, SI, TI, VAC) \tag{11}$$

The main challenge of symbolic regression is to generate the equation non-trivial that the final solution is still suitable for the new experimental data. The method to settle this problem reveals the connection among derivative of groups of variable[SL09] by comparing $\Delta x/\Delta y$ from experimental data with $\delta x/\delta y$ from a candidate equation. Schmidt defined a new metric – Partial-Derivative-Pair to measure the quality of equation, as nontrivial candidate should present $\Delta x/\Delta y \approx \delta x/\delta y$.

In each iteration, new equations are derived from recombining the previous ones and new expressions. The results of algorithm would be convergence into a group of feasible equations which have reached the predictive accuracy and complexity. The entire flow-chart of symbolic regression is shown as followed, Figure 15.

Symbolic regression is demonstrated successfully discovery physical laws with experimental data. The MOS prediction models provided by Symbolic regression are credible. However, the value of bitrate depends on the compression method and video itself. Thus, the constant of prediction model could be retained by linear regression on different dataset so that the model has a optimal performance.

Linear regression

Take an example, a possible MOS prediction model from symbolic regression is

$$MOSp = \beta_0 + \beta_1 \cdot Bitrate + \beta_2 \cdot SI + \beta_3 \cdot TI^3 + \beta_4 \cdot TI \cdot VAC + \beta_5 \cdot VAC \quad (12)$$

where $MOSp$ is the model output predicting MOS, $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]$ is the coefficients of model. Feature set of model is
$X = [Bitrate, SI, TI^3, TI \cdot VAC, VAC]$, which could be input to linear regression algorithm directly, since now $MOSp$ has a linear relationship with X.

According to *Applied regression analysis* [DSP66], linear model can be written as $MOS = X\beta + \epsilon$, where is a vector of random errors. If we have n video sequences and p features, MOS is a $n \times 1$ column, X is a $n \times (p+1)$ matrix, and β is a $(p+1) \times 1$ vector. The vector β contains unknown constants to be estimated.

To calculate the weight $\hat{\beta}$, the normal equation is

$$X^T X \hat{\beta} = X^T MOS. \quad (13)$$

Hence,

$$\hat{\beta} = (X^T X)^{-1} X^T MOS, \quad (14)$$

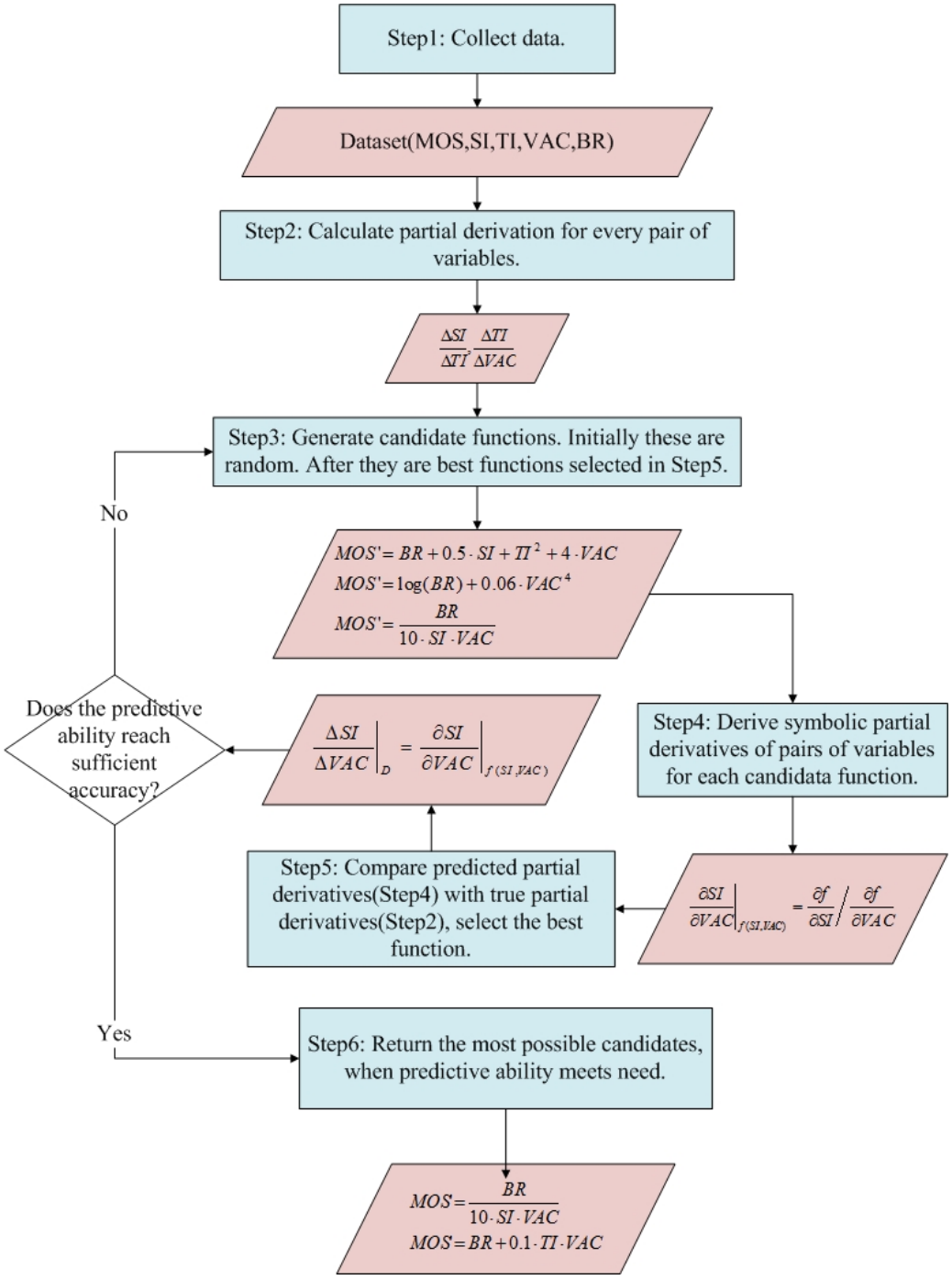


Figure 15: Flow-chart of Symbolic Regression

since the normal equation is always consistent. Then, the normal equation have a unique solution.

To be noted, if candidate formula from symbolic regression is

$$MOSp = \frac{\beta_1 Bitrate}{\beta_2 + \beta_3 Bitrate} + \frac{\beta_4 Bitrate^2 + \beta_5 Bitrate}{\beta_6 + \beta_7 Bitrate \cdot VAC^2 + \beta_8 TI^2 \cdot VAC} \quad (15)$$

, the weight vector β can be computed by nonlinear regression.

4.2.2 MOS prediction based on compression

Model 1:

$$\begin{aligned} MOSp = \beta_0 + \beta_1 Bitrate + \beta_2 Biterate^2 + \beta_3 Bitrate^3 + \beta_4 Bitrate \times VAC \\ + \beta_5 Bitrate \times VAC^2 + \beta_6 Bitrate \times VAC^3 + \beta_7 Bitrate^2 \times VAC \end{aligned} \quad (16)$$

Model 2:

$$MOSp = \frac{\beta_1 + \beta_2 Bitrate + \beta_3 TI}{\beta_4 + \beta_5 Bitrate + \beta_6 TI + \beta_7 VAC} \quad (17)$$

In these two models, feature SI has been eliminated. Actually, the performance of candidates with SI is not as good as these two. The factor leads to this occurrence is VAC represent the a kind of spatial information as it computes the video's average entropy of saliency information for each frame. Especially, in model 1, MOS are predicted by Bitrate and VAC, indicating the contribution VAC make for the subjective quality evaluation.

4.2.3 MOS prediction based on quality metric

Model 3:

$$MOSp = \beta_0 + \beta_1 PSNR + \beta_2 SI^2 + \beta_3 VAC + \beta_4 PSNR \times TI + \beta_5 SI \times VAC \quad (18)$$

Model 4:

$$MOSp = \beta_0 + \beta_1 PSNR + \beta_2 SI^2 + \beta_3 TI + \beta_4 VAC + \beta_5 SI \times VAC \quad (19)$$

4.3 Experiment and Results

4.3.1 Database

With the MOS prediction models training based on IRCCyN/IVC_SVC4QoE_QP0_QP1_Video_VGA database[PEB+11], we attempt to validate they are able to predict MOS successfully on other database. IRCCyN/IVC_Influence_Content_Video_VGA database [PBP+12], which are formed in the same condition (Table 3), except this databased is rated by 29 viewers, and same compression method (Table 2) with training database, are used in this section. The subjective video quality test was taken the methodology ACR-HR.

4.3.2 Experiment

With a view to an accurate assessment of model's performance, k-fold cross-validation are employed to the demonstration experiments. Cross-validation is used to estimate how accurately a prediction model performs in practice. To implement the cross-validation, one database is divided into complementary subsets, performing the analysis on one subset (defined as training set) and validating the analysis on the other subset (defined as validation set). In order to reduce the variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over all rounds.

In practice, there are totally 35 contents in the database, and 3 to 5 compressed sequences for each content. So, five-fold cross-validation is implemented in this experiment. In another words, in each iteration, 7 contents constitute the validation set and the rest of contents compose the training set. The whole validation experiments are run five times with each pair of training set and validation set.

4.3.3 Result and discussion

The metrics of validation set from k-fold cross-validation on four models presented in last section are shown in table. For each validation set, linear correlation coefficient(CC) and root mean square error(RMSE) are computed. Table 4 illustrates the mean and standard deviation of 5 values of CC and RMSE.

The performances among Bitrate-based model or PSNR-based model are similar. The high average CC and low average RMSE means that the quality scores from proposed models are quite close to the ground-truth MOS. The low standard deviation of CC and RMSE indicated that the model's performances on all the validation folds are steady and little effects by different contents. To draw a conclusion, they can predict the MOS with small errors by refining the linear relationship between Bitrate or PSNR and MOS with content information.

To predict quality rating based on Bitrates, Model 2 performed slightly better than Model 1. The reason could be Model 2 contains more features, which would offer more clues. Among the PSNR-based models, Model 4 achieved a few higher CC than Model 3.

	meanCC	stdCC	meanRMSE	stdRMSE
Model 1 Func.(16)	0.8454	0.0560	0.5445	0.0872
Model 2 Func.(17)	0.8538	0.0569	0.5314	0.0985
Model 3 Func.(18)	0.8636	0.0235	0.6448	0.1665
Model 4 Func.(19)	0.8641	0.0227	0.6458	0.1678
Bitrate	0.8505	/	0.9478	/
$f(Bitrate, SI, TI)$ Func.(20)	0.8278	0.0933	0.5851	0.1685
PSNR	0.8036	/	1.0530	/
$f(PSNR, RI, TI)$ Func.(21)	0.8430	0.0918	0.5964	0.1792
Video Quality Model (VQM)	0.8408	/	0.9919	/

Table 4: Results on four proposed video quality evaluation models from k-fold cross-validation

For demonstrating VAC making a contribution on MOS prediction, we directly employed Bitrate and PSNR to estimate the subject quality score. The results are shown in the table line5 and line7. After the mapping procedure, the Bitrate have a high correlation coefficient value with MOS, but it is still a little lower than Model2. Our models have a much better performance than the one only using PSNR. In addition, we trained two models with same method we proposed. These two models evaluate the MOS only with SI, TI and Bitrate or PSNR with VAC indicator, which are shown below.

$f(Bitrate, SI, TI)$:

$$\begin{aligned} MOSp = \beta_0 + \beta_1 Bitrate + \beta_2 Biterate^2 + \beta_3 Bitrate^3 \\ +\beta_4 SI + \beta_5 SI^2 + \beta_6 TI + \beta_7 SI \times TI \end{aligned} \tag{20}$$

$f(PSNR, SI, TI)$:

$$\begin{aligned} MOSp = \beta_0 + \beta_1 PSNR + \beta_2 SI + \beta_3 SI^2 + \beta_4 SI^3 \\ +\beta_5 PSNR \times SI + \beta_6 PSNR^2 \times SI + \beta_7 SI \times TI \end{aligned} \tag{21}$$

In order to validate our models with existing methods, IRCCyN/IVC_Influence_Content_Video_VGA database is evaluated by Video Quality Model (VQM) [WP]. VQM is a standardized method of objectively measuring video quality. The computed CC is 0.8408, which is lower than our four proposed models, especially Model4. The RMSE is 0.9919, much higher than the results from our models.

4.4 Conclusion

In this section, before the subjective video quality estimation with visual attention complexity has been introduced, we discussed about the subjective quality rating – Mean Opinion Score (MOS) evolution with Bitrate, PSNR and characteristic of content (i.e. VAC, SI and TI).

Afterwards, the MOS prediction models have been presented. In this paper, all models were trained with IRCCyN/IVC's H.264 compressed video database. The VAC, SI and TI are extracted from the original videos, and bitrate and PSNR are from the compressed video. To combine these features, the symbolic regression is implemented to discover the relationship between

them and MOS, finding the feasible functions for quality estimation model. Linear regression aims at optimize the coefficient of model, refining the prediction performance on each unique database. In this thesis, we offered two models respectively for Bitrate-based and PSNR-based quality estimation, which are the best possible candidates provided by symbolic regression algorithm.

In order to evaluate the MOS prediction capability, k-fold cross-validation is processed on each models. Four models have high performance on subjective quality score prediction. Last but not least, their performance is compared with other models without VAC and VQM. The proposed models both outperformed them.

5 Conclusion and Future Work

5.1 Conclusion

To draw a conclusion, firstly, this thesis contains a novel orientation in the study of visual attention complexity of scene, which is discovering the relationship between VAC and saliency information, though the proposed VAC indicator does not have a strong enough correlation with ground truth data. But, various further work could be continued in this field. Secondly, this paper generates a video quality estimation model with VAC indicator and other user-reachable features with methods of machine learning algorithm. The models were demonstrated that they are well perform on MOS prediction. By comparing our proposed models with VQM, our model outperformed VQM with higher CC (especially 0.8641), and much lower RMSE (especially 0.6458).

Moreover, to the aspect of my personal ability, my technical skills has been improved by working on this internship. Firstly, my knowledge about the human visual system, the video quality assessment and machine learning has been augmented. Secondly, my skill of the problem analyzing and solving has been improved. Last but not least, I have learned how to face the unexpected obstacles appeared during the processing.

5.2 Future Work

Certainly, there are various ways to continue the investigations started in this thesis. The future work could be on two main aspects.

For the visual attention complexity indicator, due to the limitation of entropy on saliency map, the topologies information on saliency map could be considered, which would indicate the number and position of saliency objects offering more clues for human gaze behavior while watching the video. On the other hand, Kalman Filter or Particle Filter could be used to build a model on the purpose of predicting the human eye movement types and gaze position along the video, which would be more convincible.

For the MOS prediction models, on this thesis, we only demonstrated it by correlation coefficient and compared it with VQM. To verify their perfor-

mance on the real world, they should be implemented on the compression systems, to assist the rate allocation.

References

[BCPLC09] Fadi Boulos, Wei Chen, Benoît Parrein, and Patrick Le Callet. Region-of-Interest Intra Prediction for H.264/AVC Error Resilience. In *Proceedings of IEEE International Conference on Image Processing*, pages 3109 – 3112, Cairo, Égypte, November 2009.

[CBD02] Laura Cowen, Linden Js Ball, and Judy Delin. An eye movement analysis of web page usability. In *People and Computers XVI-Memorable Yet Invisible*, pages 317–335. Springer, 2002.

[CW04] Andrea Cavallaro and Stefan Winkler. Segmentation-driven perceptual quality metrics. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 5, pages 3543–3546. IEEE, 2004.

[Dal92] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 2–15. International Society for Optics and Photonics, 1992.

[DCE11] Matthieur Perreira Da Silva, Vincent Courboulay, and Pascal Estraillier. Image complexity measure based on visual attention. *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3281 – 3284, 2011.

[DSP66] Norman Richard Draper, Harry Smith, and Elizabeth Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.

[FCLCJ12] Yunlong Feng, Gene Cheung, Patrick Le Callet, and Yusheng Ji. Video attention deviation estimation using inter-frame visual saliency map analysis. *Proc. of SPIE-IS&T Electronic Imaging Vol*, 8305:83050H, 2012.

[FCTJ11] Yunlong Feng, Gene Cheung, Wai-tian Tan, and Yusheng Ji. Hidden markov model for eye gaze prediction in networked video streaming. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[FLYW08] Xin Feng, Tao Liu, Dan Yang, and Yao Wang. Saliency based objective quality assessment of decoded video affected by packet losses. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 2560–2563. IEEE, 2008.

[FMS08] Alex Forsythe, Gerry Mulhern, and Martin Sawey. Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*, 40(1):116–129, 2008.

[For09] Alexandra Forsythe. Visual complexity: Is that all there is? In *Engineering Psychology and Cognitive Ergonomics*, pages 158–166. Springer, 2009.

[FSS03] Alex Forsythe, Noel Sheehy, and Martin Sawey. Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computers*, 35(2):334–342, 2003.

[HKP06] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[Iha93] Shunsuke Ihara. *Information theory for continuous systems*, volume 2. World Scientific, 1993.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[ITU08] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. Technical report, International Telecommunication Union, Geneva, Switzerland, April 2008.

[Koz92] John R Koza. *Genetic Programming: vol. 1, On the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[KU87] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987.

[Llo01] Seth Lloyd. Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine*, 21(4):7–8, 2001.

[Llo02] Seth Lloyd. Measures of Complexity a non–exhaustive list. 2002.

[LMNLCB10] Olivier Le Meur, Alexandre Ninassi, Patrick Le Callet, and Dominique Barba. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing Image Communication*, 25(7):547–558, 2010.

[MC09] O Le Meur and P Le Callet. What we see is most likely to be what matters: Visual attention and applications. *Proceedings of the 16th IEEE international conference on Image processing*, 2009.

[NLMLCB09] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):253–265, 2009.

[OBM98] Wilfried Osberger, Neil Bergmann, and Anthony Maeder. An automatic image quality assessment technique incorporating higher level perceptual factors. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pages 414–418. IEEE, 1998.

[OD07] Tobias Oelbaum and Klaus Diepold. A reduced reference video quality metric for avc/h. 264. In *Proc. European Signal Processing Conference*, pages 1265–1269, 2007.

[Pal99] Stephen E Palmer. *Vision science: Photons to phenomenology*. The MIT press, 1999.

[PBP+12] Yohann Pitrey, Marcus Barkowsky, Romuald Pépion, Patrick Le Callet, and Helmut Hlavacs. Influence of the source content and encoding configuration on the perceived quality for scalable video coding. In *Proceedings of the SPIE Human Vision and Electronic Imaging XVII*, volume 8291, pages 1–6, San franscisco, United States, January 2012.

[PEB+11] Yohann Pitrey, Ulrich Engelke, Marcus Barkowsky, Romuald Pépion, and Patrick Le Callet. Aligning subjective tests using a low cost common set. In *QoE for Multimedia Content Sharing*, page irccyn contribution, Lisbonne, Portugal, June 2011.

[PLCC+07] Stéphane Péchard, Patrick Le Callet, Mathieu Carnec, Dominique Barba, et al. A new methodology to estimate the impact of h. 264 artefacts on subjective video quality. In *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM2007*, 2007.

[RFS05] J Rigau, M Feixas, and M Sbert. An Information-Theoretic Framework for Image Complexity. *Computational Aesthetics in Graphics, Visualization and Imaging*, page 177, 2005.

[Roq09] LMMP Tomaz Roque. Quality evaluation of coded video. *Commands for the first pass: ffmpeg-i filename-an-pass*, pages 1–10, 2009.

[RP92] Keith Rayner and Alexander Pollatsek. Eye movements and scene perception. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):342, 1992.

[SB10] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335–350, 2010.

[SL09] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

[SV80] Joan G Snodgrass and Mary Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174, 1980.

[Tat07] Benjamin W Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[TCKW+95] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545, 1995.

[TG80] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[TR11] J. K. Tsotsos and A. Rothenstein. Computational models of visual attention. 6(1):6201, 2011.

[Win07] Stefan Winkler. Video quality and beyond. In *Proc. European Signal Processing Conference*, pages 3–7, 2007.

[Win12] Stefan Winkler. Analysis of public image and video databases for quality assessment. 2012.

[WP] Stephen Wolf and Margaret Pinson. NTIA Report 02-392: Video Quality Measurement Techniques. Technical report, Institute for Telecommunication Sciences.

[WSB03] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. Objective video quality assessment. *The handbook of video databases: design and applications*, pages 1041–1078, 2003.

[YKP10] Junyong You, Jari Korhonen, and Andrew Perkis. Attention modeling for video quality assessment: Balancing global quality and local quality. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 914–919. IEEE, 2010.